Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection*

Sricharan Kumar† Ashok Srivastava‡

Abstract

Prediction intervals provide a measure of the probable interval in which the outputs of a regression model can be expected to occur. Subsequently, these prediction intervals can be used to determine if the observed output is anomalous or not, conditioned on the input. In this paper, a procedure for determining prediction intervals for outputs of non-parametric regression models using bootstrap methods is proposed. Bootstrap methods allow for a non-parametric approach to computing prediction intervals with no specific assumptions about the sampling distribution of the noise or the data. The asymptotic fidelity of the proposed prediction intervals is theoretically proved. Subsequently, the validity of the bootstrap based prediction intervals is illustrated via simulations. Finally, the bootstrap prediction intervals are applied to the problem of anomaly detection on aviation data.

1 Introduction

In regression models, the estimated mean square error is often the only indicator of the quality of the predicted model. However, the mean square error estimate is insufficient to answer several questions about the model, including: (i) Which regions of the input space is the prediction quality good or poor?, and (ii) Conditioned on the input, which observed output values in the test set are anomalies? Such questions can be answered by specifying prediction intervals within which the predicted variable is likely to fall at a specified confidence level for a given input configuration.

1.1 Previous work In this section, we highlight previous work done in the areas of prediction intervals and anomaly detection. Subsequently, we contrast the contribution of this paper to the work highlighted in this section.

1.1.1 Prediction intervals Prediction intervals have been exclusively studied for parametric linear regression models [9]. In this body of work, two assumptions were made throughout: (i) the data was being generated via a linear model + noise, and (ii) the regression method being used was least-squares linear regression. Prediction intervals for linear models can be broadly classified into two types - (i) closed form expressions for empirical prediction intervals as a function of the data [22, 18, 19], and (ii) bootstrap based estimates of the prediction intervals [26]. The classical closed form prediction interval expressions were originally computed under the assumption of the noise being normally distributed [22], before being generalized [18, 19]. Stine [26] proposed prediction intervals based on bootstrap resampling of the training data.

1.1.2 Anomaly detection While anomaly detection has been extensively studied in literature [6], traditional anomaly detection algorithms (1-SVM [20], k-NN based methods: iOrca [2], GEM [12], BP-kNNG [24], and density based methods: LOF and iForest [16]) do not address the following problem: For multivariate data of the form $(x, y) : x \in \mathbb{R}^d, y \in \mathbb{R}$, given nominal training data and a test point (x_0, y_0), traditional anomaly detection methods are capable of checking either if the multivariate test point (x_0, y_0) is anomalous, or if the marginal test point y_0 is anomalous. However, in a scenario where the last dimension y is nominally an unknown function of the input x, traditional anomaly detection algorithms are incapable of checking if the marginal test point y_0 is nominal or anomalous *given* the input x_0.

1.2 Contribution In this paper, an algorithm for determining prediction intervals for outputs of general non-parametric regression models is proposed using bootstrap methods. In contrast to the work done in [22, 18, 19, 26, 13], where the underlying model is known to be linear, our algorithm does not make any assumptions about the underlying data model or noise distribution. Subsequently, using the prediction intervals, a novel anomaly detection algorithm is proposed to test if the observed output of a test point is an anomaly or not,

*This work is supported by the NASA Aeronautics Research Mission Directorate Seedling Program, and the NASA Aviation Safety Program, System-Wide Safety and Assurance Project.

†Research Engineer, Stinger Ghaffarian Technologies Inc. (SGT Inc.), NASA Ames Research Center, Moffett Field, CA 94035. (email:Sricharan.Kumar@nasa.gov)

‡Principal Scientist, Data Sciences group, NASA Ames Research Center, Moffett Field, CA 94035. (email:Ashok.Srivastava@nasa.gov)

conditioned on the observed input.

The rest of this paper is organized as follows. In Section 2, the problem is formally stated. In Section 3, the non-parametric regression models are described in detail. The bootstrap procedure for determining prediction intervals is described, and their validity is proved theoretically and via simulations in Section 4. The proposed theory is applied to the problem of anomaly detection in Section 5. The proposed anomaly detection algorithm is applied to determine aircraft in a fleet which are consuming abnormally large amounts of fuel. Finally, conclusions are given in Section 6.

2 Preliminaries

Throughout this paper, we assume the following model:

$$\mathbf{y}(x) = \psi(x) + \epsilon(x), \tag{2.1}$$

where $\psi(x) : \mathbb{R}^d \to \mathbb{R}$ is some unknown, but deterministic, continuous, (p, C) smooth[1] function, and $\epsilon(x) : \mathbb{R}^d \to \mathbb{R}$ is an uniform random field used to model noise. In particular, for any $x_1, .., x_t$, assume that $\epsilon(x_1), .., \epsilon(x_t)$ are iid, 0-mean and have finite variance $\sigma^2 < \infty$.

Assume that the observed data pairs (x, y) are drawn from some underlying joint density $f(x, y)$. In particular, let $\mathbf{R} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, ..\} \overset{\text{iid}}{\sim} f(x, y)$ denote a stream of observed training data. Also let $\mathbf{R}(r)$ denote the first r samples in $\mathbf{R}$: $\mathbf{R}(r) = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, .., r\}$.

2.1 Notation Denote the cumulative distribution function corresponding to $f(x, y)$ by $F(x, y)$. For some fixed set of realizations $\mathbf{R}(r)$, let $F_{\mathbf{R}(r)}(x, y)$ be the corresponding empirical distribution function [28]. Let $\mathbb{E}$ be the expectation wrt the underlying probability distribution $F(x, y)$.

Also, we will denote random variables and vectors in bold text, and indicate the weak convergence of sequences of random variables or vectors by $\Rightarrow$. Denote the l_∞ distance between two distributions $F(.), G(.)$ by

$$d_\infty(F, G) = \sup_{t \in \mathbb{R}} |F(t) - G(t)|.$$

Finally, for any given training data set $Z = \{(x_i, y_i), i = 1, .., n\}$, denote the regression model estimated using Z by $\check{y}_Z(.)$.

2.2 Problem statement Assume that r iid training realizations $\mathbf{R}(r)$ are available for estimating the model $\hat{\mathbf{y}}_r(x) = \check{y}_{\mathbf{R}(r)}(x)$ using non-parametric regression methods. Given this training data, the goal of this paper is to provide a prediction interval for a future observation $\mathbf{y}_0 = \mathbf{y}(x_0)$ corresponding to an input x_0. In particular, our goal is to determine prediction intervals at level α:

$$\mathbf{I}_{\alpha,r}(x_0) = (\mathbf{l}_{\alpha,r}(x_0), \mathbf{u}_{\alpha,r}(x_0)),$$

using the input $\mathbf{R}(r)$ such that the probability

$$Pr\{\mathbf{y}(x_0) \in \mathbf{I}_{\alpha,r}(x_0)\} \approx 1 - \alpha.$$

The last statement will be made precise in section 3.3.

3 Regression models

In this section, we discuss the estimation of the function $\psi(.)$ using non-parametric regression methods, given the set of training realizations $\mathbf{R}(r)$. Any of the popular non-parametric regression techniques including nearest neighbor regression [8], kernel regression [17], local polynomial regression [10], partitioning regression [11], support vector regression [1], artificial neural networks [23] and decision tress [5] can be used to determine the regression model. However, we require that the regression model satisfies some regularity conditions which are listed below.

3.1 Regularity assumptions on regression model The model $\hat{\mathbf{y}}_r(x)$ is assumed to be a deterministic, continuous function of the training data $\mathbf{R}(r)$. The model $\hat{\mathbf{y}}_r(x)$ is also assumed to converge to a limit denoted by $\hat{\mathbf{y}}(x)$ as $r \to \infty$. Finally, the MSE rate

$$e_r(x_0) = \mathbb{E}[\hat{\mathbf{y}}_r(x) - \psi(x)]^2$$

is assumed to decay to 0. In particular, let

$$e_r(x_0) = \mathcal{O}(r^{-2\gamma}), \tag{3.2}$$

for some $\gamma > 0$.

We note that all the non-parametric regression models listed in the previous section satisfy these assumptions. For details, please refer to [14] for nearest neighbor, polynomial and kernel regression methods, [7] for support vector regression, [15] for artificial neural networks, and [21] for decision trees.

3.1.1 Optimal rates Stone [27] showed that for (p, C) smooth function $\psi(.)$, the optimal minimax rate of convergence that can be obtained by non-parametric regression estimates is given by $r^{-2p/(2p+d)}$. Consequently, this implies that γ is bounded above by $p/(2p + d)$ and therefore by $1/2$.

[1] Please refer to Definition 1 [14] for details about (p, C) smooth functions

3.2 Bootstrapping Resample the training data $\mathbf{R}(r)$ a total of t times, each time drawing a set of r realizations with replacement. Denote these t sets of r realizations by $\mathbf{B}_i(r) = \{\mathbf{x}_{i1}, .., \mathbf{x}_{ir}\}, i = 1, .., t$. For each set of bootstrap realizations $\mathbf{B}_i(r), i = 1, .., t$, determine the regression fit $\bar{\mathbf{y}}_{i,r}(x) = \check{\mathbf{y}}_{\mathbf{B}_i(r)}(x)$. Fix $t = r^\gamma$.

3.3 Formal problem statement In the next section, we will determine prediction interval $\mathbf{I}_{\alpha,r}(x_0)$ such that

$$\lim_{r\to\infty} \frac{r^\gamma}{\sqrt{\log\log r}}[Pr\{\mathbf{y}(x_0) \in \mathbf{I}_{\alpha,r}(x_0)\} - (1-\alpha)] = 0.$$

In the sequel, denote the function $a_r(\gamma) = \sqrt{\log\log r}/r^\gamma$.

4 Prediction intervals

The output $\mathbf{y}(x_0)$ at input x_0 is forecast using $\hat{\mathbf{y}}_r(x_0)$. The prediction interval for $\mathbf{y}_0 = \mathbf{y}(x_0)$ is determined as follows. The future observation $\mathbf{y}_0$ can be written as

$$\begin{aligned} \mathbf{y}_0 = \mathbf{y}(x_0) &= \psi(x_0) + \epsilon(x_0) \\ &= \hat{\mathbf{y}}_r(x_0) + \psi(x_0) - \hat{\mathbf{y}}_r(x_0) + \epsilon(x_0) \\ &= \hat{\mathbf{y}}_r(x_0) + \eta_r(x_0) + \epsilon(x_0), \end{aligned} \tag{4.3}$$

where

$$\eta_r(x_0) = \psi(x_0) - \hat{\mathbf{y}}_r(x_0),$$

is the error in the model. To determine the prediction interval for $\mathbf{y}_0$, we first analyze the distribution of $\mathbf{y}_0 - \hat{\mathbf{y}}_r(x_0) = \eta_r(x_0) + \epsilon(x_0)$. Denote the distribution of $\epsilon(x_0)$ by $H_\epsilon(.)$ and the distribution of $\eta_r(x_0)$ by $H_{\eta_r(x_0)}(.)$.

Now, note that $\eta_r(x_0)$ and $\epsilon(x_0)$ are independent [4]. The error in the prediction centered around $\hat{\mathbf{y}}_r(x_0)$ therefore has contribution from the following two independent components: (i) the error due to the model: $\eta_r(x)$, and (ii) the error due to the observation noise: $\epsilon(x)$.

4.1 Error distribution In this section, we are going to investigate the computation of the distributions of $\eta_r(x)$ and $\epsilon(x)$.

4.1.1 Model error Realizations of the model error $\eta_r(x_0)$ can be obtained via the bootstrapping approach as follows. Let the true distribution of the oracle output $\hat{\mathbf{y}}(x_0)$ be denoted by $G(.)$ and the distribution of $\hat{\mathbf{y}}_r(x_0)$ be denoted by $G_r(.)$. Also denote the distribution of the bootstrap iid samples $\bar{\mathbf{y}}_{i,r}(x)$ by $\hat{G}_r(.)$.

A well known result of the empirical distribution function [28] is that $d_\infty(F, F_{\mathbf{R}(r)}) = \mathcal{O}(a_r(0.5))$. Then, by direct application of Lemma A.1,

$$\sup_x |\hat{G}_r(x) - G_r(x)| = \mathcal{O}(a_r(0.5)).$$

In other words, the bootstrap based samples $\bar{\mathbf{y}}_{i,r}(x_0)$ approximate the distribution of $\hat{\mathbf{y}}_r(x_0)$ up to $\mathcal{O}(a_r(0.5))$, and can therefore be used to compute quantiles of $G_r(.)$.

Let the mean of the distribution $G_r(.)$ be μ_r and the mean of $G(.)$ be μ. Then, by (3.2), $|\mu_r - \mu| = \mathcal{O}(r^{-\gamma})$. Next, let

$$\hat{\mu}_r(x_0) = \frac{\sum_{i=1}^m \bar{\mathbf{y}}_{i,r}(x_0)}{m},$$

and observe that $\hat{\mu}_r(x_0)$ is an estimate of the mean of G_r. By lemma A.3, the error $|\hat{\mu}_r(x_0) - \mu_r|$ is of order $\mathcal{O}(a_r(0.5))$ with exponentially high probability.

Denote the centered samples $\mathbf{m}_i = \bar{\mathbf{y}}_{i,r}(x_0) - \hat{\mu}_r(x_0)$, and the empirical distribution of these samples by $\hat{H}_r(.)$. Then, $\hat{H}_r(.)$ will converge weakly to the distribution of $\eta_r(x_0)$. In particular, the l_∞ distance between the two distributions, $d_\infty(H_{\eta_r(x_0)}, \hat{H}_r)$ is of order $\mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma})$.

4.1.2 Observation error Denote the differences $\mathbf{o}_i = y_i(x_i) - \hat{\mathbf{y}}_i(x_i)$. By the bias observation, we observe that these iid random variables $\mathbf{o}_i$ will converge in distribution to the true distribution of the noise $\epsilon(x)$, and the l_∞ distance between the two distributions is $\mathcal{O}(r^{-\gamma})$.

4.1.3 Distribution of total error From the previous two sections, we see that $\mathbf{m}_i, i = 1, 2, .., t$ and $\mathbf{o}_i, i = 1, 2, .., r$ correspond to samples which approximate $\eta_r(x_0)$ and $\epsilon(x_0)$ with error up to $\mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma})$.

By the independence of $\mathbf{m}_i, i = 1, 2, .., m$ and $\mathbf{o}_i, i = 1, 2, .., r$, and lemma A.2 it follows that the set $C = \{\mathbf{t}_k = \mathbf{m}_i + \mathbf{o}_j, i = 1, 2, .., m; j = 1, 2.., r\}$ corresponds to samples of the distribution $\eta_r(x_0) + \epsilon(x_0)$ with errors up to $\mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma})$.

4.2 Prediction interval The prediction interval $I_{\alpha,n}(x_0)$ can then be specified as follows. Let C_α, $C_{1-\alpha/2}$ denote the $\alpha/2$, $1-\alpha/2$ quantiles of the set C respectively. Then, define the prediction interval as

$$\mathbf{I}_{\alpha,r}(x_0) = \hat{\mathbf{y}}_r(x_0) + (C_\alpha, C_{1-\alpha/2}).$$

The procedure for computing this prediction interval is stated in the form of algorithm 1. The theoretical and experimental validity of the prediction interval are established via theorem 4.1 and in section 4.3 respectively.

THEOREM 4.1. *The prediction interval $I_{\alpha,r}(x_0)$ is an asymptotic $1-\alpha$ prediction interval in the following sense:*

$$\lim_{r\to\infty} \frac{Pr\{\mathbf{y}(x_0) \in \mathbf{I}_{\alpha,r}(x_0)\} - (1-\alpha)}{a_r(\gamma)} = 0.$$

Proof. Let the distribution of the iid realizations $\{\mathbf{t}_k\}$ be given by $\hat{T}_{s,r}$, and the distribution of $\eta_r(x_0) + \epsilon(x_0)$ be given by $\tilde{T}_{s,r}$. Then,

$$d_\infty(\tilde{T}_{s,r}, \hat{T}_{s,r}) = \mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma}).$$

Then, the quantiles C_α, $C_{1-\alpha/2}$ converge to the true quantiles of $\tilde{T}_{s,r}$ at rate $\mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma})$. This implies that $Pr\{\eta_r(x_0) + \epsilon(x_0) \in (C_\alpha, C_{1-\alpha/2})\} = 1 - \alpha + \mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma})$. Finally, observing that $\mathcal{O}(a_r(0.5)) + \mathcal{O}(r^{-\gamma}) = o(a_r(\gamma))$ concludes the proof.

Algorithm 1 Prediction intervals for regression models

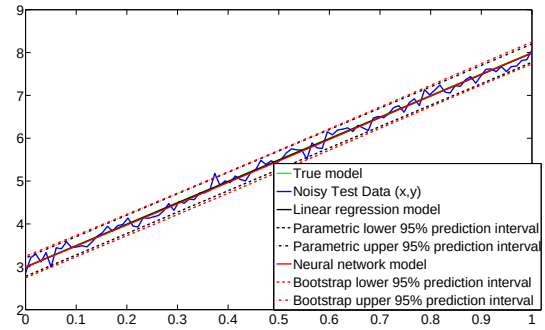
procedure PREDICTINTERVAL($R(r), \alpha, x_0$)
  Build regression model $\hat{y}_r$
  Initialize error sample set $E = \phi$
  for each training sample (x_i, y_i) **do**
    Compute error $o_i = y_i(x_i) - \hat{\mathbf{y}}_r(x_i)$
    $E \rightarrow E \cup \{o_i\}$
  end for
  Build t bootstrap samples B_i from $R(r)$
  Initialize bootstrap sample set $D = \phi$
  for each bootstrap sample B_i **do**
    Build regression models $\bar{\mathbf{y}}_{i,r}(x)$
    Obtain centered samples $\mathbf{m}_i$
    $D \rightarrow D \cup \{\mathbf{m}_i\}$
  end for
  Build the set C by convolving D and E
  Obtain C_α, $C_{1-\alpha/2}$:
    the $\alpha/2$, $1-\alpha/2$ quantiles of the set C
  Set $I_{\alpha,r}(x_0) = \hat{\mathbf{y}}_r(x_0) + \{C_\alpha, C_{1-\alpha/2}\}$
  Return $I_{\alpha,r}(x_0)$
end procedure

4.3 Simulations In this section, the proposed bootstrap based prediction intervals are validated through a series of Monte-Carlo experiments.

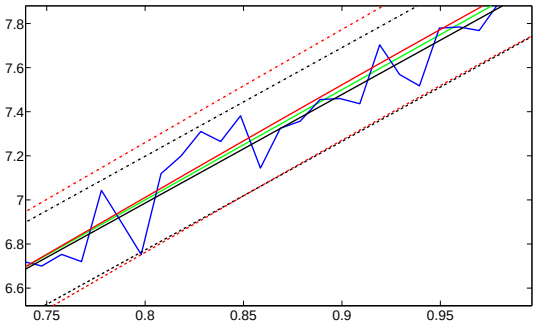
4.3.1 Simple linear model In the first experiment, we generate data drawn from a simple linear model

$$y = 3x + 5 + \epsilon,$$

where x is drawn uniformly between $[0, 1]$ and ϵ is drawn iid from $N(0, \sigma^2)$ with $\sigma = 0.1$. The size of the training set is $r = 100$. For doing regression, we use the artificial neural network regression method and then compute prediction intervals based on bootstrap samples as described in algorithm 1. Finally, we contrast the proposed prediction intervals with the classical parametric prediction intervals [22] which have been derived for linear regression.



(a)



(b)

Figure 1: Comparison of parametric 95% prediction intervals based on linear regression and and bootstrap based 95% prediction intervals based on artificial neural networks on data generated via a linear model. (b) is a zoomed version of (a). From the figures, it is clear that there is excellent agreement between the paramteric and bootstrap based prediction intervals.

In the classical model for simple linear regression, under the assumption that the noise is distributed iid $\mathcal{N}(0, \sigma^2)$, the appropriate prediction interval [22] for a future value y_0, given explanatory level x_0, is

$$J_{\alpha,n}(x_0) = \hat{y}(x_0) \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n}(\bar{x} - x_i)^2}},$$

where $\hat{\sigma}$ is the estimate of the noise.

We evaluate and plot the true response, noisy observations and 95 % prediction intervals corresponding to the parametric linear regression and non-parametric neural network regression models along $s = 100$ locations placed uniformly in the $[0, 1]$ interval in Fig. 1. From Fig. 1, it is clear that there is excellent agreement between the parametric prediction intervals based on linear regression and and bootstrap based prediction

intervals based on artificial neural networks. The prediction intervals based on the bootstrap are marginally wider, accounting for the slower $\mathcal{O}(r^{-\gamma})$ rate of convergence of the non-parametric neural network regression method in comparison to the linear regression method, which enjoys a parametric $\mathcal{O}(r^{-1/2})$ rate of convergence.

4.3.2 Simple polynomial model To illustrate the non-parametric advantage that our bootstrap prediction interval algorithm enjoys, we repeat the previous experiment, but with one difference: we generate the training data according to the model:

$$y = 3x^2 + 5 + \epsilon.$$

The results are shown in Fig. 2. From Fig. 2, it is clear that the parametric prediction interval is inaccurate due to mis-specification of the model. On the other hand, the non-parametric bootstrap based prediction interval is able to accurately determine the correct prediction interval.

4.3.3 Multivariate example In the next experiment, we consider the following non-linear model

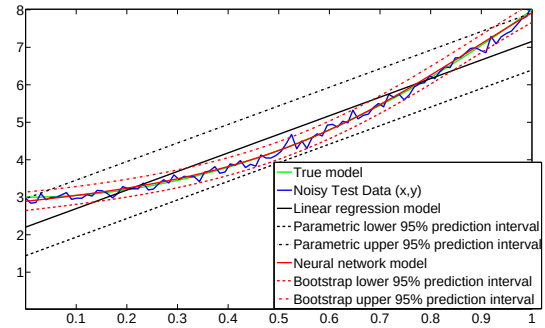
$$y = \exp(x_1) + x_2 x_3^2 + \log(|x_4 + x_5|) + \epsilon,$$

with $x = (x_1, .., x_5)$ drawn from a multivariate random variable with mean vector $m = [0.1, 0.2, 0, 0.05, 1.2]$ and covariance matrix

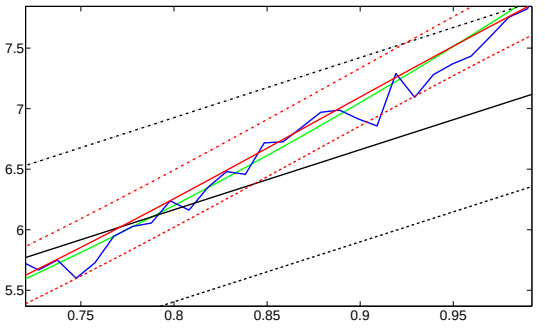
$$\Sigma = \begin{bmatrix} 1.00 & 0.43 & 0.45 & -0.29 & -0.69 \\ 0.43 & 1.00 & 0.25 & -0.36 & -0.36 \\ 0.45 & -0.36 & 1.00 & -0.91 & -0.36 \\ -0.29 & -0.36 & -0.91 & 1.00 & 0.49 \\ -0.69 & -0.36 & -0.36 & 0.49 & 1.00 \end{bmatrix}$$

We draw a total of $r = 1000$ training samples, and determine the prediction intervals corresponding to $\alpha = \{.005, .01, .02, .05, .1, .2\}$ at $s = 100$ uniformly placed points in the $[0, 1]^5$ grid. Subsequently, we compute the experimentally observed coverage width of the prediction intervals for each of the desired coverage widths $\alpha \in \{.005, .01, .02, .05, .1, .2\}$. The results are shown in table 1. From table 1, it is clear that there is excellent agreement between the two sets of values.

For visual illustration, we plot the true response, the noisy test samples, and the estimated prediction intervals as a function of the sorted index of the true response in Fig. 3. From the figure, it is clear that the estimated prediction intervals accurately capture the variation in the observed test data. In particular, we note that the prediction intervals are thinner at locations where the variation in y is smaller and vice versa. This is in complete agreement with our intuition.



(a)



(b)

Figure 2: Comparison of parametric 95% prediction intervals based on linear regression and and bootstrap based 95% prediction intervals based on artificial neural networks on data generated via a quadratic model. (b) is a zoomed version of (a). From the figures, it is clear that the parametric prediction interval is inaccurate due to mis-specification of the model. On the other hand, the non-parametric bootstrap based prediction interval is able to accurately determine the correct prediction interval.

Desired and observed coverage						
Desired	.20	.10	.05	.02	.01	.005
Observed	.198	.112	.042	.019	.012	.006

Table 1: Desired and observed coverage of the proposed prediction intervals. There is excellent agreement between the two sets of values.

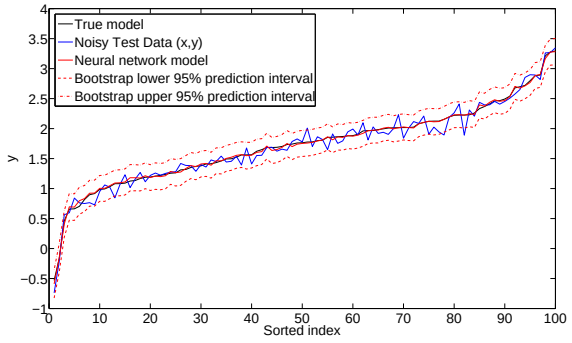


Figure 3: Prediction intervals for non-linear model. The estimated prediction intervals accurately capture the variation in the observed test data.

5 Application to anomaly detection

Prediction intervals can be used to determine anomalies in the test data in the following manner. Given the training data $\mathbf{R}(r)$, a test point (x_0, y_0), and a desired false alarm rate α, we construct the prediction interval $\mathbf{I}_{\alpha,r}(x_0)$ using Algorithm 1. We then declare (x_0, y_0) to be an anomaly at false alarm rate α if $y_0 \notin \mathbf{I}_{\alpha,r}(x_0)$ and nominal otherwise. This Conditional Anomaly Detection (CAD) scheme for testing a set of anomalies $S(s) = (x_i, y_i), i = 1, .., s$ is formally described in Algorithm 2.

Algorithm 2 Conditional Anomaly Detection

procedure DETECTANOMALY($\mathbf{R}(r), S(s), \alpha$)
  Initialize anomaly set $A = \phi$
  for each test sample (x_i, y_i) in test set $S(s)$ **do**
    Build prediction interval: $\mathbf{I}_{\alpha,r}(x_i)$:
    $\mathbf{I}_{\alpha,r}(x_i) = \text{PredictInterval}(\mathbf{R}(r), \alpha, x_i)$
    if $(y_i \notin \mathbf{I}_{\alpha,r}(x_i))$ **then**
      $A \rightarrow A \cup (x_i, y_i)$
    end if
  end for
  Return A
end procedure

5.1 Comparison to existing methods

The proposed CAD algorithm differs from popular anomaly detection algorithms (1-SVM [20], k-NN based methods: iOrca, GEM, BP-kNNG, and density based methods: LOF and iForest) in the following respect. Traditional anomaly detection methods check for anomalies by testing for the following null and alternate hypothesis [12]:

$$H0 : (x_0, y_0) \sim f(x, y) \quad \text{versus} \quad H1 : (x_0, y_0) \not\sim f(x, y).$$

If the data of interest is only the y variable, the anomaly detection algorithms could be used to test between the hypothesis:

$$H0 : y_0 \sim f(y) \quad \text{versus} \quad H1 : y_0 \not\sim f(y).$$

In contrast, the proposed anomaly detection method is checking for anomalies by testing between:

$$H0 : y_0 \sim f(y|x_0) \quad \text{versus} \quad H1 : y_0 \not\sim f(y|x_0).$$

In other words, the traditional algorithms test for anomalies wrt the joint distribution $f(x, y)$ or the marginal distribution $f(y)$ of y , whereas the proposed algorithm tests for anomalies wrt the conditional distribution $f(y|x)$.

5.2 Aviation fuel study

The environmental impact of aviation is enormous given the fact that in the US alone there are nearly 6 million flights per year of commercial aircraft. Detecting aircraft in a fleet which consume excess fuel is therefore extremely important. The proposed CAD algorithm can be used to determine, on a on a second by second basis, the nominal fuel consumption interval as a function of the measured state - velocity, acceleration etc - of the aircraft. The prediction intervals can then be used to determine if the instantaneous fuel consumption of the flight is anomalous.

This fuel study differs from the current state-of-the-art used by airline companies, which involves simply comparing the actual fuel consumption against averages for a given flight or aircraft. Such computations do not sufficiently control for the context of the flight and therefore may not reveal more subtle performance issues.

5.2.1 Data set

The data set used in our study is known as Flight Operational Quality Assurance (FOQA), which is used for numerous purposes including improving safety and efficiency of the operations of commercial and business transport aircraft. In addition to the actual fuel consumption, FOQA data includes parameters such as velocity, acceleration, pitch rate, payload etc. This data is measured and monitored on every

aircraft on a second-by-second basis. The full list of parameters can be found in Table 1 [25].

In this study, we use FOQA data from an airline company corresponding to the one year period between May 2010 - April 2011. This data set has FOQA information of about 60,000 flights, corresponding to 330 distinct aircraft. The first 6 months of this data is used for training, and the next 6 months are used as the test data set $S(s)$. The full description of these data sets can be found in [25].

5.2.2 Application of CAD The FOQA parameters (suitably normalized) are treated as inputs, and the observed fuel consumption rate values are treated as outputs. This training data is fed as input to CAD, and is used to check for anomalies in the later half of the year. The results obtained for each time point are aggregated on a per-aircraft basis. The results corresponding to the sorted top $k = 5$ aircraft wrt % anomaly time are shown in table 2. From the table, it is clear that only two aircraft are detected to be consuming excess fuel. Furthermore, these anomalies aircraft were not detected when using a traditional anomaly detection method like iOrca.

Aircraft*	# of flights	% anomaly time
147	68	32.43%
641	111	14.51%
111	104	0.85%
976	45	0.37%
342	33	0.37%
...	...	...

Table 2: List of aircraft ($*$ = anonymized) and the % anomaly time as determined by CAD. Only two aircraft in the fleet of 330 aircraft were found to be anomalous.

5.2.3 Results On further investigation, it was found that the longitudinal acceleration sensors were faulty in both aircraft. In both cases, the sensor measurements were lower than the true acceleration of the aircraft. This resulted in prediction intervals which were shifted down, and subsequently resulted in the true fuel consumption falling outside the prediction intervals and leading to these aircraft being classified as anomalous. This has since been officially confirmed by the airline company, and the faulty FOQA sensors in these aircraft have since been replaced.

6 Conclusions

In this paper, a procedure for determining prediction intervals in non-parametric regression models for a future observation is specified. The procedure is based on bootstrap techniques and does not require any underlying assumptions about the data or noise model. The validity of the bootstrap based prediction intervals is shown in theory to hold asymptotically. Subsequently, the validity of the intervals is proved via monte-carlo experiments.

Next, an anomaly detection algorithm based on prediction intervals is proposed. The anomaly detection algorithm differs from popular anomaly detection algorithms in that it discovers anomalies wrt conditional probability distributions. The anomaly detection algorithm is applied to discover aircraft in fleet which consume excess fuel.

A General results

Consider a probability space $(\Omega, \mathbb{B}, \mathbb{P})$. Let $\mathbf{X}_k : \Omega \to \mathbb{R}$, $k = 1, 2, ..$ be a sequence of random variables with corresponding distribution functions $F_k(.)$. Also, assume that this sequence converges in distribution to a random variable $\mathbf{X} : \Omega \to \mathbb{R}$ with distribution function $F(.)$. In particular, let

$$d_\infty(F_k, F) = \mathcal{O}(k^{-\beta}),$$

for some $\beta > 0$.

Identically define a sequence of random variables $\mathbf{Y}_k : \Omega \to \mathbb{R}$, $k = 1, 2, ..$ with distribution functions $G_k(.)$ and a weak convergence limit to random variable $\mathbf{Y}$ with distribution G. Similarly, let

$$d_\infty(G_k, G) = \mathcal{O}(k^{-\gamma}),$$

for some $\gamma > 0$.

Construct sequences of random variables corresponding to $F_k(.)$, $F(.)$ as follows. Let $\bar{\mathbf{X}}_k = \{\mathbf{X}_{k1}, \mathbf{X}_{k2}, ..\}$ where the elements of $\bar{\mathbf{X}}_k$ are drawn iid from $F_k(.)$. Define $\bar{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2, ..\}$ in an identical manner with elements drawn from $F(.)$. Also let $\bar{\mathbf{X}}_k(r) = \{\mathbf{X}_{k1}, \mathbf{X}_{k2}, .., \mathbf{X}_{kr}\}$ and $\bar{\mathbf{X}}(r) = \{\mathbf{X}_1, \mathbf{X}_2, .., \mathbf{X}_r\}$ be subsets of size r. Let $X = \{x_1, x_2, ...\}$ be any fixed sequence of real numbers, and let $X(r) = \{x_1, .., x_r\}$ denote a subset of size r.

Finally, define a sequence of functions $h_i : \mathbb{R}^i \to \mathbb{R}, i = 1, 2, ...$ Assume that this sequence of function have a limit in the following sense: for any $\epsilon > 0$ and any sequence X, there exists $r_0(\epsilon) \in \mathbb{N}$ such that for all $r \geq r_0$,

$$|h_{r+1}(X(r+1)) - h_r(X(r))| < \epsilon.$$

Define the limit of this sequence by $h : \mathbb{R}^\infty \to \mathbb{R} = \lim_{r\to\infty} h_r(X(r))$. Also, define the sets $H_i(\alpha) = \{X(i) : h_i(X(i)) \leq \alpha\}$ for some $\alpha \in \mathbb{R}$. Now, we will prove the lemma:

LEMMA A.1. *For any $r \in \mathbb{N}$, $h_r(\bar{\mathbf{X}}_k)$ converges weakly to $h_r(\bar{\mathbf{X}})$. Furthermore, the error rate*

$$\sup_\alpha |Pr\{h_r(\bar{\mathbf{X}}_k) \leq \alpha\} - Pr\{h_r(\bar{\mathbf{X}}) \leq \alpha\}| = \mathcal{O}(k^{-\beta}).$$

Proof. By the Cramer-Wold device [3] and the fact that the distributions $F_k(.)$ converge to $F(.)$, it clearly follows that $\bar{\mathbf{X}}_k(r)$ converges weakly to $\bar{\mathbf{X}}(r)$ for all $r \in \mathbb{R}$. By the continuous mapping theorem, this in turn implies that $h_r(\bar{\mathbf{X}}_k)$ converges weakly to $h_r(\bar{\mathbf{X}})$.

To prove the next part of the statement, consider:

$$\begin{aligned} Pr\{h_r(\bar{\mathbf{X}}_k) \leq \alpha\} &= Pr\{\bar{\mathbf{X}}_k(r) \in H_r(\alpha)\} \\ &\quad (\text{since } d_\infty(F_k, F) = \mathcal{O}(k^{-\beta})) \\ &= Pr\{\bar{\mathbf{X}}(r) \in H_r(\alpha)\} + \mathcal{O}(k^{-\beta}) \\ &= Pr\{h_r(\bar{\mathbf{X}}(r)) \leq \alpha\} + \mathcal{O}(k^{-\beta}). \end{aligned}$$

Observing that this is true for any $\alpha \in \mathbb{R}$ concludes the proof.

LEMMA A.2. *The sequence of random variables $\mathbf{X}_k + \mathbf{Y}_k$ converge weakly to $\mathbf{X} + \mathbf{Y}$. Furthermore, if $\mathbf{X}_k, \mathbf{Y}_k$ are independent for every k, the distribution $H_k(.)$ of $\mathbf{X}_k + \mathbf{Y}_k$ converges to the distribution $H(.)$ of $\mathbf{X} + \mathbf{Y}$ at the rate:*

$$d_\infty(H_k, H) = \mathcal{O}(k^{-\min\{\beta,\gamma\}}).$$

Proof. By the Cramer-Wold device [3], it is trivial to see that the joint pairs $(\mathbf{X}_k, \mathbf{Y}_k)$ converge weakly to $(\mathbf{X}, \mathbf{Y})$. The result follows by invoking the continuous mapping theorem [3].

$$\begin{aligned} Pr\{\mathbf{X}_k + \mathbf{Y}_k \leq \alpha\} &= \int_{\alpha_1} F_k(\alpha - \alpha_1) dG_k(\alpha_1) \\ &= \int_{\alpha_1} F(\alpha - \alpha_1) dG(\alpha_1) \\ &\quad + \mathcal{O}(k^{-\beta}) + \mathcal{O}(k^{-\gamma}) \\ &= \int_{\alpha_1} F(\alpha - \alpha_1) dG(\alpha_1) \\ &\quad + \mathcal{O}(k^{-\min\beta,\gamma}) \\ &= Pr\{\mathbf{X} + \mathbf{Y} \leq \alpha\} \\ &\quad + \mathcal{O}(k^{-\min\{\beta,\gamma\}}). \end{aligned}$$

LEMMA A.3. *Denote the sample mean of $\bar{\mathbf{X}}(r)$ by μ_r and the mean of $F(.)$ by μ. For any $\gamma < 1/2$, denote the event*

$$E_\gamma = I\{|\mu_r - \mu| \geq \log\log r / r^\gamma\}.$$

Then,

$$Pr\{E_\gamma\} \leq \exp(-r^{(1/2-\gamma)}).$$

Proof. The statement follows by direct application of the Chernoff bound.

References

[1] D. Basak, S. Pal, and D. Patranabis, *Support vector regression*, Neural Information Processing–Letters and Reviews, 11 (2007), pp. 203–224.

[2] K. Bhaduri, B. Matthews, and C. Giannella, *Algorithms for speeding up distance-based outlier detection*, in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 859–867.

[3] P. Billingsley, *Convergence of probability measures*, vol. 493, Wiley-Interscience, 2009.

[4] L. Breiman, *The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error*, Journal of the American Statistical Association, 87 (1992), pp. 738–754.

[5] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and regression trees*, Chapman & Hall/CRC, 1984.

[6] V. Chandola, A. Banerjee, and V. Kumar, *Anomaly detection: A survey*, ACM Computing Surveys (CSUR), 41 (2009), p. 15.

[7] A. Christmann, A. Van Messem, and I. Steinwart, *On consistency and robustness properties of support vector machines for heavy-tailed distributions*, Statistics and Its Interface, 2 (2009), pp. 311–327.

[8] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi, *On the strong universal consistency of nearest neighbor regression function estimates*, The Annals of Statistics, 22 (1994), pp. 1371–1385.

[9] N. Draper and H. Smith, *Applied regression analysis new york, 1998*.

[10] J. Fan, I. Gijbels, T. Hu, and L. Huang, *An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation*, Statistica Sinica, 6 (1996), pp. 113–127.

[11] S. Guthery, *Partition regression*, Journal of the American Statistical Association, 69 (1974), pp. 945–947.

[12] A. Hero III, *Geometric entropy minimization (gem) for anomaly detection and localization*, Ann Arbor, 1001 (2006), pp. 48109–2122.

[13] J. Hwang and A. Ding, *Prediction intervals for artificial neural networks*, Journal of the American Statistical Association, 92 (1997), pp. 748–757.

[14] M. Kohler, A. Krzyżak, and H. Walk, *Optimal global rates of convergence for nonparametric regression with unbounded data*, Journal of Statistical Planning and Inference, 139 (2009), pp. 1286–1296.

[15] M. Kohler and J. Mehnert, *Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors*, Neural Networks, 24 (2011), pp. 273–279.

[16] F. Liu, K. Ting, and Z. Zhou, *Isolation forest*, in Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE, 2008, pp. 413–422.

[17] Y. Mack and B. Silverman, *Weak and strong uniform consistency of kernel regression estimates*, Probability Theory and Related Fields, 61 (1982), pp. 405–415.

[18] D. Olive, *Prediction intervals for regression models*, Computational statistics & data analysis, 51 (2007), pp. 3115–3122.

[19] R. Schmoyer, *Asymptotically valid prediction intervals for linear models*, Technometrics, 34 (1992), pp. 399–408.

[20] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, *Estimating the support of a high-dimensional distribution*, Neural computation, 13 (2001), pp. 1443–1471.

[21] C. Scott and R. Nowak, *Minimax-optimal classification with dyadic decision trees*, Information Theory, IEEE Transactions on, 52 (2006), pp. 1335–1353.

[22] G. Seber and A. Lee, *Linear regression analysis*, vol. 936, Wiley, 2012.

[23] D. Specht, *A general regression neural network*, Neural Networks, IEEE Transactions on, 2 (1991), pp. 568–576.

[24] K. Sricharan and A. Hero III, *Efficient anomaly detection using bipartite k-nn graphs*, Ann Arbor, 1001, p. 48104.

[25] A. Srivastava, *Greener aviation with virtual sensors: a case study*, Data Mining and Knowledge Discovery, (2012), pp. 1–29.

[26] R. Stine, *Bootstrap prediction intervals for regression*, Journal of the American Statistical Association, (1985), pp. 1026–1031.

[27] C. Stone, *Optimal global rates of convergence for nonparametric regression*, The Annals of Statistics, 10 (1982), pp. 1040–1053.

[28] A. Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.